RELATÓRIO TÉCNICO: IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO K-NEAREST NEIGHBORS (KNN) APLICADO AO INSTAGRAM

Nome dos Residentes:
Mirella Oliveira Rebouças
Quele da Silva Andrade

Feira de Santana, Bahia
17 de dezembro de 2024

Resumo

Este relatório apresenta a implementação do algoritmo k-Nearest Neighbors (kNN) para categorizar influenciadores do Instagram com base em métricas como seguidores, taxa de engajamento e média de curtidas. O projeto inclui análise exploratória, otimização de hiperparâmetros e avaliação do modelo, destacando seus pontos fortes e limitações. Apesar de resultados satisfatórios para classes majoritárias, o desbalanceamento dos dados comprometeu o desempenho geral, ressaltando a necessidade de abordagens complementares.

1. Introdução

A popularidade das redes sociais tornou os influenciadores digitais atores centrais em estratégias de marketing. Nesse contexto, surge a necessidade de analisar e categorizar influenciadores com base em métricas quantitativas. Vrontis et al. (2021), em sua revisão sistemática de 155 artigos, destacam a evolução do marketing de influência nas redes sociais e propõem um framework integrado que enfatiza a importância de métricas como número de seguidores e taxa de engajamento para a análise efetiva de influenciadores digitais.

Este projeto utiliza o algoritmo kNN devido à sua simplicidade e adequação para problemas de classificação, explorando sua aplicabilidade em dados reais e complexos do Instagram. Como fundamentado por Zafarani et al. (2014) em seu trabalho seminal sobre mineração de dados em mídias sociais, a escolha de algoritmos adequados e o pré-processamento correto dos dados são cruciais para análises efetivas em redes sociais. O objetivo principal foi analisar um conjunto de dados de influenciadores do Instagram, contendo informações como:

- Seguidores: Quantidade total de seguidores do influenciador.
- Média de curtidas: Média de curtidas nos últimos posts.
- Taxa de engajamento: Engajamento percentual dos últimos 60 dias.
- Total de curtidas: Soma total das curtidas nos posts.

Morton (2020) enfatiza em seu estudo exploratório sobre motivações de usuários jovens a importância de compreender métricas quantitativas para avaliar o impacto real dos influenciadores. O autor destaca que métricas como taxa de engajamento e número de seguidores são indicadores cruciais para entender a efetividade de um influenciador digital.

Desafios encontrados incluíram valores ausentes, representação inconsistente de variáveis (e.g., "3.3k", "2.9m"), e forte desbalanceamento entre as classes. Abaixo na Tabela 1 tem o resumo do *dataset*.

Tabela 1. Resumo do *Dataset*

Coluna	Tipo	Exemplos
followers	*float64*	6.5m, 5.9m

avg_likes	*object*	3.3k, 2.9k
60_day_eng_rate	*object*	1.39%, 2.23%
total_likes	*float64*	665.3k, 3.3m

Fonte: própria autora.

2. Metodologia

2.1 Análise Exploratória

- Identificação de colunas categóricas com valores como "3.3k" ou "5.2m".
- Conversão de valores para formato numérico, utilizando multiplicadores (k = mil; m = milhão).
- Verificação de variáveis nulas e preenchimento ou exclusão conforme necessário.

A Tabela 2, apresentada a seguir, detalha de forma mais aprofundada os problemas identificados e as soluções implementadas durante a análise exploratória dos dados.

Tabela 2. Problemas Identificados e Soluções Implementadas na Análise Exploratória de Dados

Coluna	Tipo inicial	Problemas Encontrados	Soluções Aplicadas
followers	*object*	Valores em k e m	Conversão para *float*
avg_likes	*object*	Valores ausentes	Preenchimento com média
60_day_eng_rate	*object*	Percentuais em formato %	Conversão para *float*

Fonte: própria autora.

Transformações de Dados Realizadas:

1. Conversão de variáveis para formato numérico.
2. Normalização para garantir escala uniforme entre as variáveis.
3. Categorização da variável *country* em continentes.

2.2 Implementação do kNN

O algoritmo kNN foi configurado inicialmente com os seguintes parâmetros:

- Número de vizinhos (k): 5.
- Métrica de distância: *Euclidiana*.
- Validação cruzada: 5-*fold*.

Durante a otimização, valores alternativos para k (1 a 15) e métricas (*Manhattan* e *Minkowski*) foram testados.

2.3 Validação e Ajuste de Hiperparâmetros

Realizada validação cruzada com *GridSearchCV* para otimizar:

- Número de vizinhos (k).
- Métrica de distância (*Euclidiana, Manhattan*).

Tabela 3. Parâmetros Testados e seus valores

Parâmetros Testados	Valores
Número de Vizinhos	3, 5, 7, 9
Métrica de Distância	*Euclidiana* e *Manhattan*

Fonte: própria autora.

Melhor configuração encontrada:

- k = 3.
- Distância *Manhattan.*

3. Resultados

Os resultados apresentados na Tabela 4 de métricas de desempenho mostram que o desempenho do modelo k-Nearest Neighbors (kNN) foi altamente impactado pelo desbalanceamento das classes no conjunto de dados. A classe 23, que possui maior representatividade (13 de 25 exemplos), apresentou valores significativos nas métricas de *precision* (0.52), *recall* (1.00) e *f1-score* (0.68). Entretanto, as demais classes, representadas por poucos exemplos no conjunto de dados, como 2, 4, 5, 8, 10, 11, 17 e 22, tiveram todas as métricas iguais a 0. Isso evidencia que o modelo está aprendendo exclusivamente a prever a classe majoritária, negligenciando completamente as classes minoritárias.

As métricas de desempenho são apresentadas abaixo:

Tabela 4: Desempenho Apresentado

Classe	*Precision*	*Recall*	*F1-Score*	Suporte
2	0,00	0,00	0,00	3
4	0,00	0,00	0,00	1
5	0,00	0,00	0,00	1
8	0,00	0,00	0,00	1
23	0,52	1,00	0,68	13
Média	0,06	0,11	0,08	25

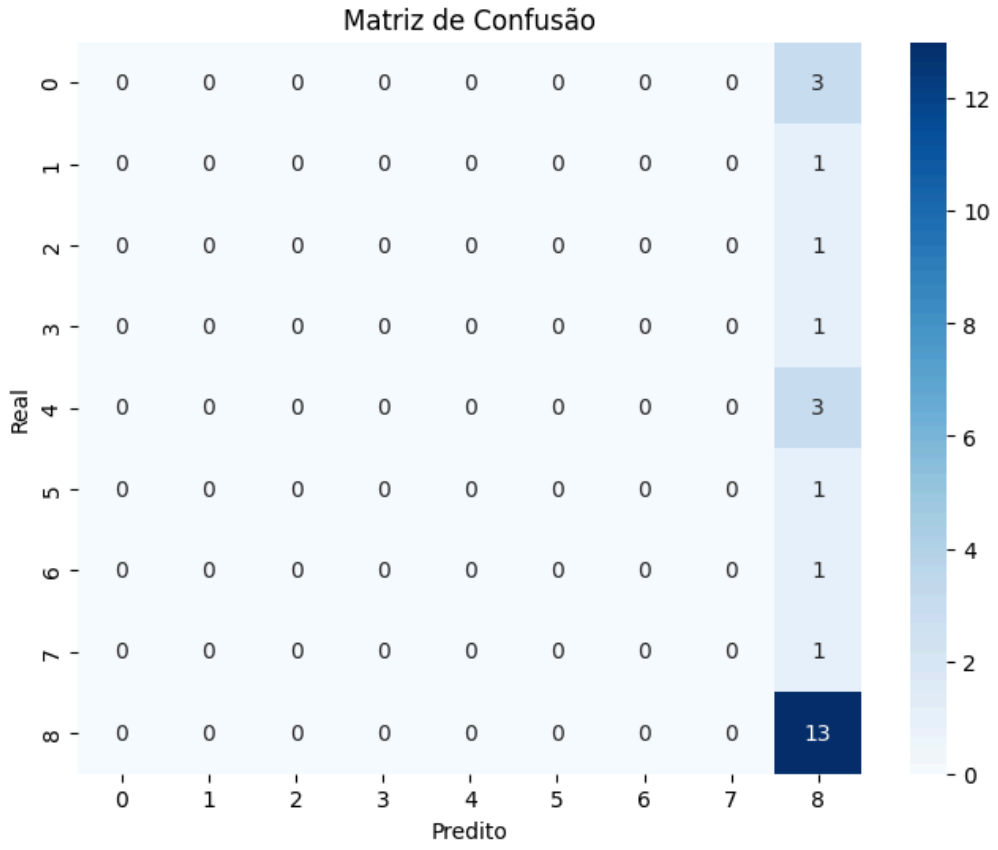
Fonte: própria autora.

Esse comportamento do modelo pode ser atribuído a características inerentes ao algoritmo kNN. Como ele baseia suas previsões na proximidade entre os dados no espaço de atributos, classes minoritárias tendem a ser sub-representadas, dificultando sua correta classificação. Além disso, as variáveis do conjunto de dados, como *followers*, que foi convertida de sufixos como "k" e "m" para valores numéricos, e outras métricas com escalas diferentes, podem ter prejudicado o cálculo de distâncias no espaço multidimensional.

Entre as principais limitações identificadas, destaca-se o desbalanceamento

do conjunto de dados. A predominância da classe 23 levou o modelo a favorecer essa classe, resultando em zero recall para as demais, o que significa que o modelo não conseguiu identificar nenhuma amostra pertencente a essas classes. Outro ponto importante é a representação dos dados, em que variáveis com escalas diferentes, como percentuais (*60_day_eng_rate*) e valores absolutos (*avg_likes*), podem ter contribuído para a má performance. Além disso, a sensibilidade do kNN à escala e ao balanceamento dos dados tornou o modelo inadequado para este cenário específico sem o uso de técnicas complementares. Abaixo a matriz de confusão com o número de previsões corretas e incorretas feitas pelo modelo.

Figura 1. Matriz de Confusão



Para superar essas limitações e melhorar os resultados, algumas soluções podem ser implementadas. Primeiramente, é essencial balancear o conjunto de dados. Isso pode ser feito por meio de técnicas como oversampling, que aumenta o número de amostras das classes minoritárias usando métodos como *SMOTE* (*Synthetic Minority Oversampling Technique*), ou *undersampling*, que reduz o número de amostras da classe majoritária. Outra estratégia é adotar a

reamostragem estratificada durante a divisão dos dados em treinamento e teste, garantindo que cada conjunto tenha uma proporção representativa de todas as classes.

Além disso, a engenharia de features desempenha um papel crucial. Padronizar todas as variáveis para evitar que diferenças de escala prejudiquem o cálculo de distâncias é fundamental. Também pode-se criar novas variáveis que combinem métricas existentes, como a razão entre seguidores e média de curtidas, para capturar padrões mais representativos. Outra abordagem seria utilizar métricas de avaliação mais adequadas para cenários desbalanceados, como o F1-macro, que dá igual peso a todas as classes, independentemente de sua representatividade.

Considerando o desempenho do kNN, pode ser interessante explorar algoritmos alternativos mais robustos ao desbalanceamento de classes. Modelos baseados em árvores, como Random Forest ou Gradient Boosting, são mais adequados para lidar com esse tipo de problema sem a necessidade de transformar o conjunto de dados. Outra opção seria o uso de *Support Vector Machines* (SVM) com kernel, que é eficaz para separar classes não linearmente separáveis, aumentando a generalização.

Por fim, para garantir que o modelo seja avaliado de forma justa durante a otimização, é recomendada a utilização de validação cruzada estratificada, que preserva as proporções de cada classe em cada *fold*. A implementação dessas estratégias pode melhorar o aprendizado do modelo em relação às classes minoritárias, aumentando a *recall* e a *precision* dessas classes. Isso também contribuirá para previsões mais confiáveis em um cenário desbalanceado, equilibrando o impacto das classes majoritárias e minoritárias, e permitindo uma análise mais justa e precisa de influenciadores no Instagram.

4. Discussão

O desempenho do modelo kNN foi diretamente impactado pela distribuição desbalanceada das classes no conjunto de dados. O modelo apresentou bom desempenho apenas para a classe majoritária (23), com métricas como recall de 1.00 e f1-score de 0.68, enquanto todas as outras classes obtiveram métricas iguais a zero. Esse resultado evidencia que o modelo aprendeu a classificar quase exclusivamente a classe majoritária, falhando em generalizar para as classes minoritárias. Essa limitação é inerente ao kNN, especialmente em cenários com desbalanceamento extremo, onde classes menos representadas são praticamente ignoradas devido à falta de vizinhos próximos suficientes para influenciar a classificação.

A sensibilidade do kNN à escala das variáveis também pode ter prejudicado a performance do modelo. No pré-processamento, foi realizada a conversão de variáveis que continham sufixos como "k" e "m" em valores numéricos, garantindo consistência no formato dos dados. No entanto, sem padronizar as variáveis, aquelas com valores em escalas maiores, como o número de seguidores, podem ter dominado o cálculo de distâncias, ofuscando informações relevantes de variáveis em escalas menores, como taxas percentuais.

O uso de validação cruzada foi uma escolha acertada para garantir uma avaliação robusta do modelo. Entretanto, a ausência de estratégias específicas para lidar com o desbalanceamento, como validação cruzada estratificada, comprometeu a capacidade de medir de forma justa o desempenho do modelo em todas as classes. Além disso, o critério de avaliação usado, baseado na acurácia, pode ser enganoso em cenários com desbalanceamento extremo, pois favorece a classe majoritária.

5. Conclusão e Trabalhos Futuros

Este projeto demonstrou a aplicabilidade do kNN na análise de influenciadores do Instagram, mas também destacou suas limitações. Entre os principais aprendizados estão:

- A importância de tratar variáveis não numéricas e desbalanceamento.
- A necessidade de ajustes cuidadosos nos hiperparâmetros do modelo.

Trabalhos futuros devem focar em:

- Adoção de técnicas de balanceamento de dados.
- Exploração de algoritmos alternativos como SVM ou modelos baseados em árvores.
- Ampliação do conjunto de dados para incluir mais classes representativas.

6. Referências

VRONTIS, D.; MAKRIDES, A.; CHRISTOFI, M.; THRASSOU, A. Social media influencer marketing: A systematic review, integrative framework and future research agenda. International Journal of Consumer Studies, v. 45, 2021. DOI: 10.1111/ijcs.12647.

ZAFARANI, R.; ABBASI, M. A.; LIU, H. Social Media Mining: An Introduction. Cambridge: Cambridge University Press, 2014. ISBN: 9781139088510.

MORTON, F. Influencer marketing: An exploratory study on the motivations of young users to follow social media influencers. Journal of Digital & Social Media Marketing, v. 8, n. 4, p. 285-295, 2020. DOI: 10.1362/204440820X15633628993371.